BigData@Chalmers
Machine Learning
Business Intelligence, Culturomics and Life Sciences

Devdatt Dubhashi
LAB
(Machine Learning, Algorithms, Computational Biology)
D&IT Chalmers
Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is
doing it, so everyone claims they
are doing it...

(Dan Ariely)
Inside the Web Intelligence Machine

“Kuo expects that Apple will introduce an iPhone 5S around June or July of this year”

“...opposition organizers plan to meet on Thursday to protest...”

640,000 Real-time Sources

Structured Data Elements
- Linguistic Processing
- Scores
- Events
- Expert Opinion

Entities

Time

Events

Recorded Future Index

8+ Billion Time-tagged Facts

Recorded Future Premium Online Analysis

Recorded Future API Customer Analytic Environment

www.recordedfuture.com
Entity Disambiguation

- Match names in text with the entity behind them

- Fundamental problem, addressed at annual competitions like Semeval

- Disambiguation is needed everywhere. Databases, web mining, linguistics, …

- Used at Recorded Future (exemplified next!)
Chris Anderson
No, His Name Is Not Ted
Chris Anderson, Curator of TED Talks, Builds his Brand

By DAVID HOCHMAN | MARCH 7, 2014

Chris Anderson was sitting in a very low-power pose. Off to the side at a staff meeting at TED's Hudson Street headquarters in January, he was folded forward with his hand on his neck, a posture that communicates self-protection according to the 2012 TED Talk on body language by the social psychologist Amy Cuddy (15.7 million views).

By letting his employees give mini
Chris Anderson,

No, His Name Is Not Ted

Chris Anderson, Curator of TED Talks, Builds his Brand

Playing career

Anderson

Anderson

Anderson

Anderson

Anderson
Entity disambiguation
“Judge a man by the company he keeps.”

- Euripides
Chris Anderson was born in Pakistan in 1957, one of three children. His parents were medical missionaries, and he spent most of his early life in Pakistan, India, and Afghanistan. He studied at a boarding school in the Himalayan mountains of India, Woodstock School, before moving to a boarding school in Bath, UK, and Oxford to study Physics. Then, he changed to Politics, Philosophy and Economics. Eventually graduating in 1980, he launched a publishing company devoted initially to hobbyist amateur magazines. In 1994, he left to launch magazines including Business 2.0 with his wife Jacqueline Novogratz to discuss topics such as pan-Arabism and the Arab Spring. Anderson is a frequent speaker at TED conferences, and has personally curated more than 120 talks. In 2014, he was named one of the 100 Most Influential People by Time magazine. In June 2011, the viewing figure stood at more than five million for a single week. The TED Prize was awarded in 2009 to Chris Anderson and three other TED Fellows, allowing hundreds of independently organised events about the power of visual media at TED Global.
Graph Communities

“Chris Anderson”

TED

WIRED
Classification with Graph Kernels
Graph Embeddings and Kernels

- Embed discrete combinatorial object (graph) into continuous Euclidean space
- Define kernel based on geometry of Euclidean sp.
- V. Jethava et al. NIPS 2012, JMLR 2013
- T. Kerola, L. Hermansson, V. Jethava, F. Johansson CIKM 2013
Demonstrator at Recorded Future

- Classifies names as *ambiguous* or *unique*

- Uses *graph classification* to classify *occurrence graphs* of names

- Achieved *state-of-the-art* results (CIKM, 2013).

- Powerful extension for complete disambiguation in progress …

- Parallel/Distributed implementation in *GraphLab*
Towards a knowledge-based culturomics

- **Språkbanken** (Swedish Language Bank), University of Gothenburg
- **Language Technology**, Lund University
- **LAB** Group Department of Computer Science and Engineering, Chalmers University of Technology
Word Embeddings

The use of word representations has become a key “secret sauce” in the success of many NLP systems in recent years, across tasks including entity recognition, part-of-speech tagging, parsing, and semantic role labeling. (Luong et al. (2013))
Deep Learning (Neural Networks)

- Revolutionized vision and speech systems
  - Dramatic improvements in image classification – near human level.
  - Skype real time translation from English to Chinese.
Word Embeddings capture meaning

\[ W(\text{"woman"}) - W(\text{"man"}) \approx W(\text{"aunt"}) - W(\text{"uncle"}) \]

\[ W(\text{"woman"}) - W(\text{"man"}) \approx W(\text{"queen"}) - W(\text{"king"}) \]
Dealing with information overload

GABA from reactive astrocytes impairs memory in mouse models of Alzheimer's disease

Clusterin contributes to caspase-3-independent brain injury following neonatal hypoxia-ischemia

Directly influences cell death is lacking. Following neonatal H-I brain injury in mice (a model of neuronal caspase-3 activation), clusterin had 50% less brain injury. Clusterin had no effect on caspase-3 activation, did not colocalize to the same sites, and did not affect exogenous purified astrocyte-secreted necrotic death. These results suggest that clusterin modulates non-caspase-dependent aspects of injury.
Document summarization

Word vectors
+ Multiple
Kernel
learning
+ Submodular
optimization
M. Kågeback, O. Mogren et al,
“Extractive Summarization
using Continuous Vector
Space Models”,
Workshop on
(CVSC) EACL 2014
The report said Andreas Lubitz repeatedly set the plane for an unauthorised descent earlier that day. He had locked the flight captain out of the cockpit. Five minutes on the Duesseldorf-Barcelona flight 07:21:10 - Plane told to descend to 21,000ft.

The co-pilot of the Germanwings plane that crashed in the French Alps in March appears to have practised a rapid descent on a previous flight, a report by French investigators says. Lubitz is suspected of deliberately crashing the Airbus 320, killing all 150 people on board. [He had locked the flight captain out of the cockpit.]
Geology
Name the three types of rock.

1. Classic
2. Punk
3. Hard
M. Kageback, F. Johansson et al, “Neural context embeddings for automatic discovery of word senses”, (NAACL 2015 workshop on Vector Space Modeling for NLP)

Used an innovative clustering technique
Exploited word and context vectors.
Senses of for ‘paper’

Vis using t-sne
Entity Linking

1. Encode entity context as set of vectors
2. Compare vector set to Wikipedia pages using point set kernel
Entity Linking

3.
Link entity to strongest match
Probabilistic Regulation of Metabolism

- Prediction of metabolic changes due to genetic or environmental perturbations
- Diagnosing metabolic disorders
- Discovering novel drug targets.
Genetic Regulation of Metabolism: Using Factor Graphs and Belief Propagation
genetic regulatory network consisting of transcription factor genes, target genes and metabolic reactions
Privacy

- Data mining with Differential Privacy
- Programming language technology for differential privacy (Sands)
- Privacy policies for social networks (Schneider)
April 14: Large-scale inference and decision making

- 8:00 opening
- 9:00 Marc Deisenroth, Imperial College, UK: Gaussian processes and big data problems.
- 11:00 Mattias Villani, Linköping University, Sweden: Efficient data subsampling strategies for speeding up MCMC.
- 13:30 Olivier Teystaud, INRIA, France: Power Network Optimisation
- 15:30 Josephine Sullivan, KTH, Sweden: Large scale ConvNets for computer vision.

April 15: Learning in biology and medicine

- 9:00 Opening
- 10:00 Devdatt Dubhashi, Chalmers University of Technology, Sweden: Network analysis in computational biology.
- 13:00 Tom Heskes, Radboud University, Nijmegen, Netherlands: Approximate message passing and biomedicine.
- 15:00 Lars Carlsson, Astra Zeneca, Sweden: Conformal prediction and drug design.

April 16: Sequential Inference and Decision making

- 09:00 Opening
- 10:00 Thomas Schön, Uppsala University, Sweden: Solving inference problems using sequential Monte Carlo
- 13:00 Ronald Ortner, University of Leoben, Austria: Regret bounds for multi-armed bandits and reinforcement learning in MDPs.
- 15:00 Christos Dimitrakakis, Chalmers University of Technology, Sweden: Approximate Bayesian Computation
Big Data Analytics May 25-29

- Hadoop
- Spark
- Spotfire

CHALMERS
• SVMs and Kernel Methods
• Graph Theoretic Methods
• Probabilistic Graphical Models
• Deep Learning
• Bayesian Decision Theory
• Reinforcement Learning

• Business intelligence
• Natural Language Technology
• Life Sciences
• Transport (Volvo)
• Infectious disease epidemiology
• Medical Imaging
• Political Science …
Chalmers Data-X?

• Life Science and Engineering
• Transport
• Energy
• Smart Cities (Built Environment)
• Production

• Volvo cars (connected cars, historical data)
• AstraZeneca (mining medical literature)
• Seal software (mining legal contracts)
Data Science vs EScience

- Data-centric
- Probabilistic models
- GPUs
- Computational biology, NLP, social sciences …

- Computation-centric
- Simulation
- Large clusters/grids
- Physics, Turbulent flows, Climate …